33.3 Via-Switch FPGA: 65nm CMOS Implementation and Architecture Extension for AI Applications

Masanori Hashimoto[1], Xu Bai[2], Naoki Banno[2], Munehiro Tada[2], Toshitsugu Sakamoto[2], Jaehoon Yu[1], Ryutaro Doi[1], Yusuke Araki[3], Hidetoshi Onodera[3], Takashi Imagawa[4], Hiroyuki Ochi[4], Kazutoshi Wakabayashi[5], Yukio Mitsuyama[6], Tadahiko Sugibayashi[2]

[1]Osaka University, Suita, Japan, [2]NEC, Tsukuba, Japan
[3]Kyoto University, Kyoto, Japan, [4]Ritsumeikan University, Kusatsu, Japan
[5]NEC, Kawasaki, Japan, [6]Kochi University of Technology, Kami, Japan

FPGAs are a suitable platform for implementing up-to-date machine learning algorithms and state-of-the-art AI applications including inference engines in embedded systems and training accelerators in cloud systems. Despite its short design turn-around time, the achievable performance is limited by the low area efficiency originating from field programmability [1-2]. Also, data transfer minimization in both amount and distance is essential for higher energy efficiency, but conventional FPGAs often require pipeline registers at SRAM and DSP I/Os to conceal long communication latency originating from non-uniform tile architecture. In pursuit of an energy-efficient FPGA platform for AI applications, a via-switch FPGA (VS-FPGA), whose programmability is attained by non-volatile via-switch crossbars in BEOL, has been proposed with the aim of utilizing FEOL fully for computing [3], but its silicon implementation is not presented yet. This work demonstrates the first implementation of VS-FPGA in 65nm CMOS and further demonstrates an AI-oriented FPGA architecture.

Figure 33.3.1 shows the key features of VS-FPGA. Non-volatile via-switch in BEOL is responsible for configuration, and hence computing solely exploits FEOL while most of FEOL is consumed by programmability in conventional SRAM FPGA. As a result, VS-FPGA improves area efficiency leading to higher performance and energy efficiency.

Figure 33.3.2 illustrates the structure and usage of via-switch in VS-FPGA. A via-switch consists of two atom switches in series and two varistors [4]. Two control lines connected to two varistors enable multiple fanouts in a via-switch crossbar, which makes signal routing flexible. Each atom switch is a non-volatile and rewritable solid-electrolyte switch, and it is composed of a solid-electrolyte sandwiched between Cu and Ru electrodes [5]. By applying a positive or negative voltage, a Cu bridge can be formed or removed. The varistor is a non-linear selector device that provides current in programming mode and isolates the control line from the signal line in normal operation mode. Configurable logic block (CLB) includes a via-switch crossbar serving as a connection block (CB) for logic block and a switch block (SB) for routing. There are also via-switches between the crossbars in adjacent CLBs for enabling and disenabling inter-CLB connections. These inter-CLB via-switches are off during crossbar programming, and they are programmed after the crossbar programming.

Figure 33.3.3 shows the die micrograph, TEM images and specifications of a chip fabricated in 65nm CMOS, where FEOL and M1-M4 are fabricated in a commercial fab and via-switch, M5 and M6-M7 (semi-global) are processed by ourselves. The die includes 6×6 CLBs and peripheral drivers in 293 × 395 μm^2, and each CLB consumes 35.55×30.7 μm^2. Note that the area of peripheral drivers is negligible for larger CLB arrays. A truth table in each LUT is implemented with another via-switch crossbar. Via-switch including atom switches, varistors, control and signal lines in Fig. 33.3.2 uses M4 and M5 layers and occupies $48F^2$ whereas its footprint can be reduced to $18F^2$ if four metal layers are used for via-switch implementation [3]. The two signal lines between which two atom switches exist are in M4 layers, and the two control lines are in M5 layers above the varistors.

Figure 33.3.4 demonstrates the area reduction thanks to via-switch. CLBs with the same functionality, each of which includes two 4-input LUTs, CB and SB, are laid out in 65nm technology. Atom switch FPGA (AS-FPGA) [6] uses two atom switches and a transistor as a cross-point switch, and both FEOL and BEOL are used. On the other hand, thanks to using BEOL varistors as selector devices in this work, the CLB area is reduced by 61.4%. Compared with the area of SRAM-type CLB reported in [6], the area is reduced to 8.3%.

Figure 33.3.5 shows a Shmoo plot of an area-minimized 16-bit counter mapped on the fabricated VS-FPGA with via-switch programming. At 1.2V, 83MHz operation is confirmed. The 0.675V limit comes from DFF operation in CLB. The resistance of via-switch in normal mode is independent of supply voltage, and it never limits Vmin. Figure 33.3.5 also shows the speed improvement referring to the Shmoo plot of AS-FPGA in [6]. VS-FPGA provides a better trade-off between clock frequency and operating voltage. VS-FPGA achieves 50MHz at 0.9V while AS-FPGA needs 0.95V for the same frequency.

Figure 33.3.6 illustrates an FPGA architecture that exploits via-switch crossbars for AI applications. There are two types of CLB, which are SRAM_CLB and Arith_CLB, and they are uniformly tiled for achieving local data movement between memories and arithmetic units. Via-switch consumes no FEOL, and hence SRAM and arithmetic circuits are packed under crossbars. For efficient systolic array implementation, local data track between adjacent Arith_CLBs is equipped in FEOL layer. These features eliminate long-distance communication and enable near-memory computing for higher energy efficiency and smaller latency. LUT is mainly responsible for FSM implementation. In addition to popular arithmetic operations, such as MAC, in AI applications, word-size multiplexers can be implemented in Arith_CLB, which reduces LUT usage and improves compatibility with high-level synthesis. We implemented the SRAM_CLB and Arith_CLB with Verilog and laid them out with commercial physical synthesis flow using several custom cells in 65nm technology. The areas of SRAM_CLB with a 90×123 crossbar and Arith_CLB with a 90×127 crossbar are 199.8 × 140.4 μm^2 and 118.8 × 140.4 μm^2, respectively. To balance BEOL crossbar area and FEOL arithmetic circuit area, two Arith_CLB contains one arithmetic circuit block. Note that the SRAM macro provided from fab uses M4, and the number of via-switches on the SRAM is limited. When SRAM macro is designed only with M1-M3, the width of SRAM_CLB can be reduced to 171.0μm.

Figure 33.3.7 shows the estimated computation density and energy efficiency. For comparison, we also implemented multiplexer-based SRAM_CLB and Arith_CLB composing a MUX-FPGA with Verilog using commercial physical synthesis flow. Supposing tensor multiplication on the FPGA with 1,152 CLBs, where the critical path consists of 5-stage LUT path in control logic, we estimated the speed and power using HSPICE simulation with a transistor-level netlist including parasitic resistance and capacitance of wires and via-switches. SRAM access power is estimated with reading and writing frequencies obtained from logic simulation results of tensor multiplication. VS-FPGA attains 29× and 5× higher computation density and energy efficiency, respectively, compared to MUX-FPGA. We also scaled the 65nm design to a commercial 28nm and ASAP 7nm [7] technologies. The on-state resistance of a via-switch is supposed to be constant since the Cu bridge size is below 10nm and smaller than F size even in 7nm technology. Other characteristics are scaled according to F. The computation density and energy efficiency are expected to improve according to technology advancement.

In conclusion, we have demonstrated the first implementation of an area and power-optimized FPGA that utilizes via-switches as programmable cross-points in a crossbar and presented a near-memory computing-oriented VS-FPGA architecture for AI applications.

Acknowledgement:
This work is supported by JST CREST under Grant JPMJCR1432.

References:
[1] I. Kuon et al., "Measuring the Gap Between FPGAs and ASICs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 2, pp. 203–215, Feb. 2007.
[2] M. Lin et al., "Performance Benefits of Monolithically Stacked 3-D FPGA," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 2, pp. 216–229, Feb. 2007.
[3] H. Ochi et al., "Via-Switch FPGA: Highly-Dense Mixed-Grained Reconfigurable Architecture with Overlay Via-Switch Crossbars," *IEEE Transactions on VLSI Systems*, vol. 26, no. 12, pp. 2723-2736, Dec. 2018.
[4] N. Banno et al., "Low-Power Crossbar Switch with Two-Varistors Selected Complementary Atom Switch (2V-1CAS; Via-Switch) for Nonvolatile FPGA," *IEEE Transactions on Electron Devices*, vol. 66, no. 8, pp. 3331-3336, Aug. 2019.
[5] M. Tada et al., "Polymer Solid-Electrolyte Switch Embedded on CMOS for Nonvolatile Crossbar Switch," *IEEE Transactions on Electron Devices*, vol. 58, no. 12, pp. 4398–4405, Dec. 2011.
[6] M. Miyamura et al., "First Demonstration of Logic Mapping on Nonvolatile Programmable Cell Using Complementary Atom Switch," *IEDM Technical Digest*, pp. 10.6.1–10.6.4, Dec. 2012.
[7] L. T. Clark et al., "ASAP7: A 7-nm finFET Predictive Process Design Kit," *Microelectronics Journal*, vol. 53, pp. 105 -115, July 2016.

 978-1-7281-3205-1/20/$31.00

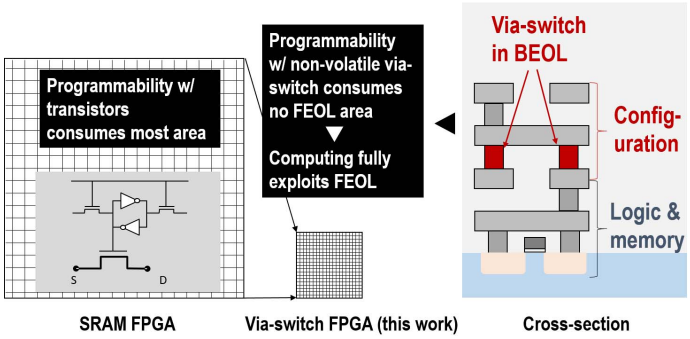


Figure 33.3.1: Background of via-switch FPGA development achieving high area efficiency.

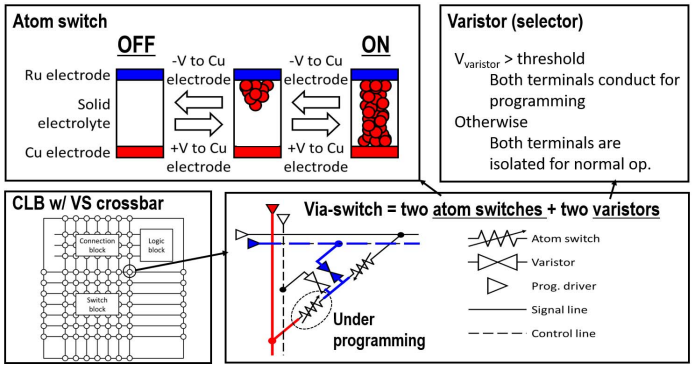
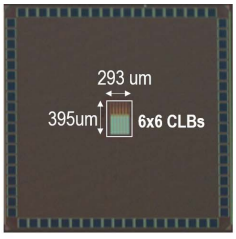
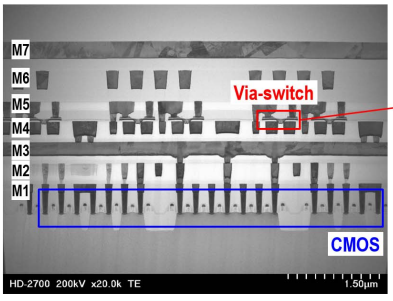


Figure 33.3.2: Structure of via-switch crossbar.



Technology	65-nm CMOS (for FEOL and M1-M4) + Post process (for via-switch and M5-M7)
# of CLBs	6 x 6
One CLB area	35.55 x 30.7um^2



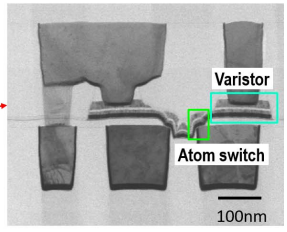


Figure 33.3.3: Die micrograph, TEM images, and specification of VS-FPGA.

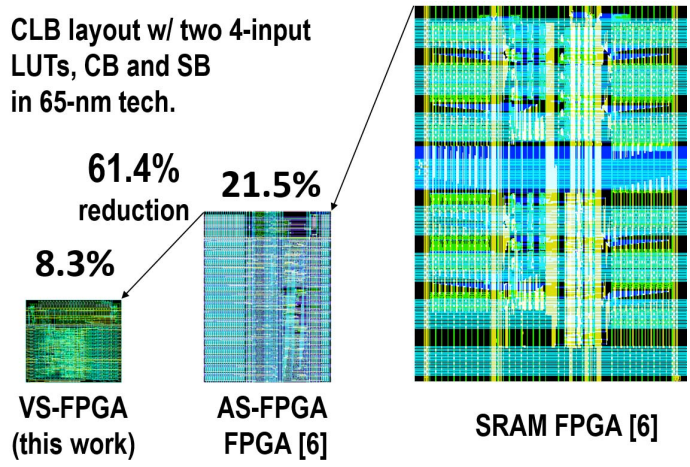


Figure 33.3.4: CLB (configurable logic block) area comparison between SRAM FPGA, AS-FPGA, and VS-FPGA. A CLB includes two 4-input LUTs, a CB (connection block) and an SB (switch block).

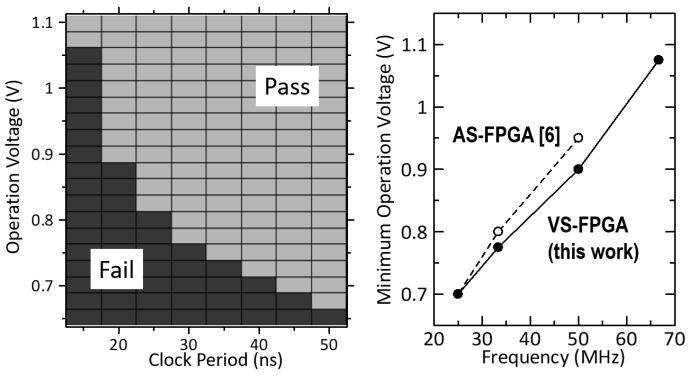


Figure 33.3.5: Measured Shmoo plot of a 16-bit counter mapped on a fabricated VS-FPGA, and comparison with AS-FPGA.

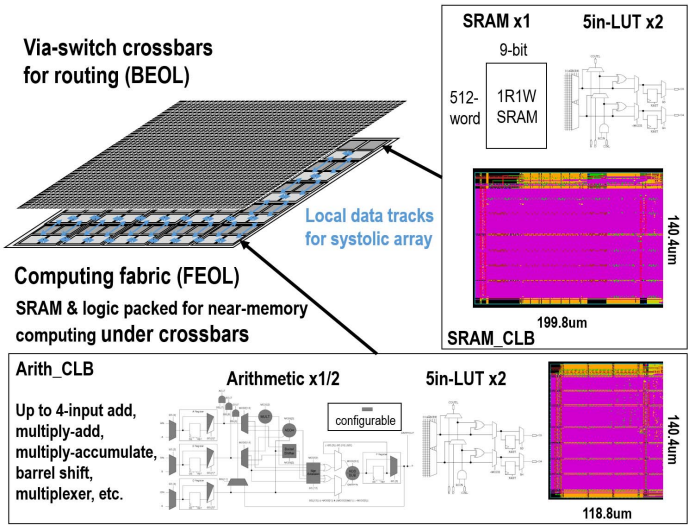


Figure 33.3.6: VS-FPGA architecture extended for near-memory computing aiming at AI applications.

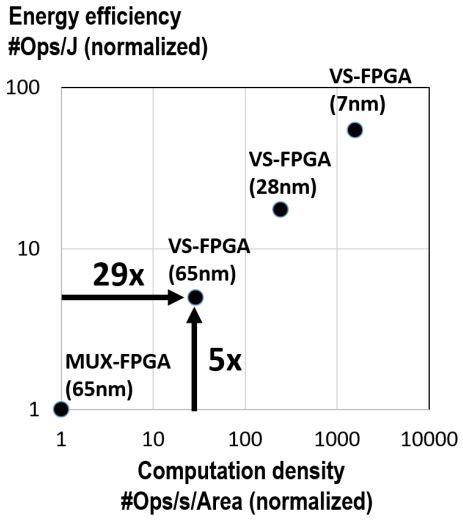


Figure 33.3.7: Comparison in computation density and energy efficiency between 65nm MUX-FPGA, 65nm VS-FPGA, 28nm VS-FPGA, and 7nm VS-FPGA.

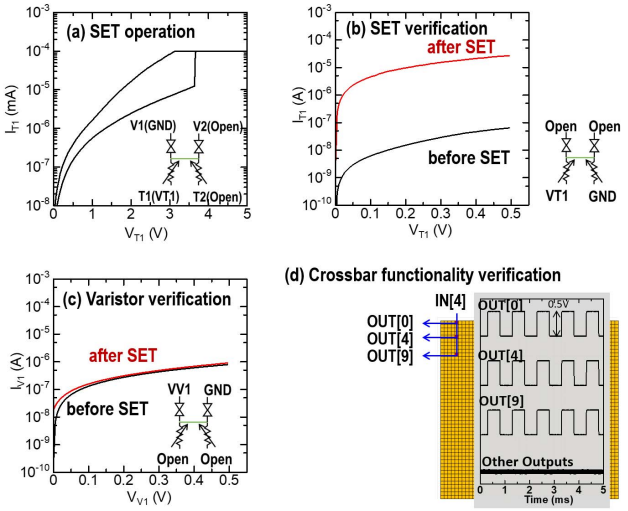


Figure 33.3.S1: Measurement results. (a) Expected hysteresis after SET operation. (b) Open and short states. (c) Varistor is functioning even after SET operation. (d) Multi-fanout routing in a fabricated 50×50 crossbar.

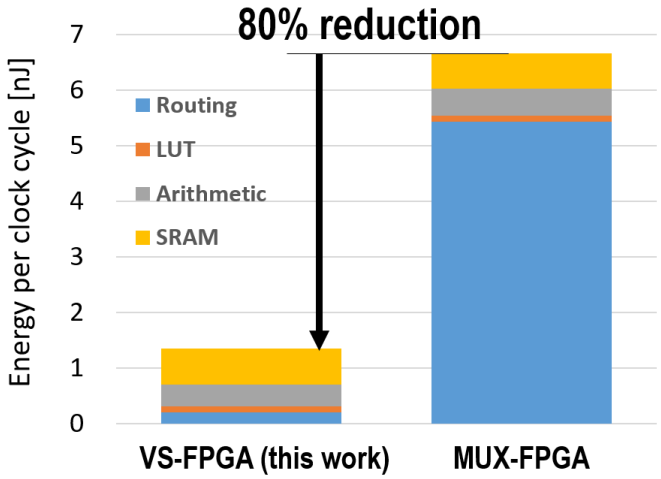


Figure 33.3.S2: Breakdown of energy consumed in VS-FPGA and MUX-FPGA in 65nm CMOS. In MUX-FPGA, 82% of energy is consumed for routing while its ratio decreases to 20% in VS-FPGA. The overall energy reduction is 80%.

978-1-7281-3205-1/20/$31.00 ©2020 IEEE